Data-driven identification of fixed expressions and their modifiability

Villada Moirón, María Begoña

Chapter 3

Automatic extraction methods

3.1 Introduction

We start with a broad overview of two different approaches to identification of collocations in corpora: the purely statistical approach and the hybrid approach. Both types of approaches employ probabilistic statistics; they mainly differ in the preparation of the extraction data. To avoid redundant description, the remaining sections in this chapter describe the different tasks in the identification process. When appropriate we include references to the purely statistical approach, or else we limit the presentation to hybrid approaches.

3.1.1 Overview of identification models

Early models were built to identify those word combinations showing strong lexical affinities in textual corpora; these lexical affinities were modelled as a strong statistical dependence between the component words (Damerau, 1993; Dunning, 1993; Pedersen et al., 1996; Pedersen, 1996). With improvements in part-of-speech tagging, phrase chunking and parsing, data which was automatically annotated with linguistic information became available; these developments in natural language processing enabled identification models that extract only those collocation candidates that satisfied certain morpho-syntactic requirements (verb–object, adj–noun, verb–PP inside the same predicate clause, etc.). Following Krenn (2000b), we refer to these models as hybrid models because they make use of both linguistic information and statistics. Currently, more sophisticated models aim at identifying multi-word lexemes and classifying them; such methods incorporate not only morpho-syntactic annotation in the extraction data, but they also use semantic ontologies such as an automatically constructed thesaurus (Lin, 1998)

or WordNet (Schone and Jurafsky, 2001; Baldwin et al., 2003; McCarthy et al., 2003) during the classification task.[1]

3.1.2 Purely statistical approach

Built on the assumption that co-occurrence frequency is a good indicator of collocativity, this approach is based on Firth's notion of collocation:

> 'Collocations of a given word are statements of the habitual or customary places of that word ... The collocation of a word or a 'piece' is not to be regarded as mere juxtaposition, it is an order of *mutual expectancy* (Firth, 1957, p. 181).'

Firth's definition stresses a relation of habitual co-occurrence and mutual expectancy between the component words of a collocation. If two words occur together a lot, then that is evidence that there is a special relation between them. A relation of mutual expectancy means that each component word in a collocation expects the other component word(s) in its surrounding context.

Purely statistical approaches apply *ngram models*.[2] Candidates (ngrams or word tuples) in datasets represent word combinations extracted from raw corpora. Ngrams may correspond to sequences of adjacent or non-adjacent words. To allow for the fact that component words in collocations may be separate from each other a numeric span is used. A window technique is used to extract ngrams of non-adjacent words observed within the boundaries of a specified span. For example, if each word w_i in a corpus is tallied with every word occurring within a window of 10 words, word w_i is combined with every word occurring from five words before w_i to five words after w_i.

To avoid prohibitively expensive computational effort, some models use a *seed word* list while sampling outcomes (Weeber et al., 2000). Such a list includes those words whose collocates are being sought. For example, if we are interested in finding collocations such as *pilot study*, the noun *study* is entered in the seed word list. Seed words are typically used together with a numeric span in which case the seed word is tallied with each co-occurring word within a specific span. An alternative heuristic is to use *stop lists* that include function words (determiners, prepositions, etc.) and/or diacritics. The words in the stop lists are considered unlikely to participate in interesting lexical associations. Damerau (1993), for example, discards all word pairs 'containing one of a defined list of common words'. The use of seed

[1] Chapter 7 briefly characterizes this research.

[2] An ngram is a sequence of n words.

words and stop lists reduces computational work and discards uninteresting candidates.

Outcomes in the dataset are ranked on the basis of a score; this score typically reflects the statistical dependence between the individual words inside the word tuple. The simplest method sorts candidate outcomes according to their raw frequency. One expects that the most frequent word combinations are better collocations than the less frequent ones. Alternatively, commonly used statistical measures are: mutual information (Church and Hanks, 1990), 'dice' coefficient and log-likelihood (Dunning, 1993), among others.

One advantage of a purely statistical approach pertains to the dataset extraction task: the sampling process is straightforward; one only needs to collect all possible tuples from a text (their length typically ranges from 2 to 5). Even though tuple extraction is easy, there is free software that collects all possible tuples from an input text. An example is the *ngram* statistics package (Banerjee and Pedersen, 2003).[3] Another advantage is that (in theory) there is no restriction on the type of collocations; thus, two word, three word, or n word collocations, where the word may belong to any part-of-speech category, could be identified. In practice, two-word collocations are the most common target due to the difficulty in adjusting standard statistical tests to handle tuples with $n \geq 3$. An exception is *mutual expectation*, a measure proposed in Dias et al. (1999).

The first disadvantage of a purely statistical approach (already suggested) is the difficulty of adjusting statistical measures to compute the association score for ngrams (tuples) with $n \geq 3$. Many statistical measures are not easily generalized to ngrams (Dias et al., 1999; Blaheta and Johnson, 2001). A second disadvantage concerns the amount of noise in the extracted lists. Often 'adverbial, adjectival, prepositional and conjunctive locutions' are also retrieved (Dias et al., 1999). If no filtering techniques are used, candidates with high scores often contain a function word next to a content word (noun, adjective, verb); these word combinations may be very frequent in a corpus but they only qualify as collocations under a purely Firthian definition.[4] To improve performance of a purely statistical approach, Weeber et al. (2000) claim that considering only window-size as a parameter is not enough to reach an optimal recall in the extraction task. The reason is that tests such as log-likelihood or Fisher's test reach several high peaks (high recall) at different window-sizes. The windowing technique is unreliable for outcomes with a frequency of less than 5 (Weeber et al., 2000). A final drawback – but

[3] *ngram* is a SourceForge project available at `http://sourceforge.net/projects/ngram`.

[4] In section 3.1.3, we introduce a linguistic notion of collocation.

not specific to this approach – is related to evaluation. Evaluation ought to reflect the accuracy with which the chosen statistics identify true collocations and multi-word lexemes in the input text. Typically, collocation extraction aims at expanding and improving existing lexica, thus the lack of reference data is always a problem. This makes evaluation rather difficult, especially when the list of potential collocations includes expressions of various sizes and linguistic types (functional, semantic and uncharacterizable ones (Daille, 1996)). Often, a human evaluator needs to assess the validity of the retrieved expressions.

Purely statistical approaches have been mainly applied to restricted domains. Dias et al. (1999) apply their method to the identification of multi-word lexical units in the European Parlament corpora and Weeber et al. (2000) investigate MEDLINE, a database of medical abstracts. The retrieved lists of collocations, terms and multi-word lexemes may be good enough for some applications such as information extraction, information retrieval and (limited) machine translation. However, other applications such as parsing and grammar development require a detailed (morpho-syntactic) characterization and classification of the multi-word lexemes potentially found in general domain text. This is one reason to investigate hybrid approaches.

3.1.3 Hybrid approaches: linguistics and statistics

Hybrid approaches assume a more restricted notion of collocation, one that we will refer to as the 'linguistic notion'. These approaches to collocation identification attempt to capture arbitrary word combinations whose component words share a syntagmatic relationship. We chose a definition proposed by Fontenelle (1992):

> 'The term *collocation* refers to the idiosyncratic syntagmatic combination of lexical items and is independent of word class or syntactic structure (Fontenelle, 1992, p. 222)'

Hybrid approaches combine linguistic information with statistics. During pre-processing, linguistic information is added to the extraction corpora. The linguistic annotation allows the researcher to specify (morpho-)syntactic patterns that describe the collocation type(s) to be extracted. Thus, the sampling process involves the search for all instances of a pre-specified template or pattern in the extraction data. Frequent patterns are adjective-noun, verb-object, verb-prepositional phrase and verb-particle combinations.

Researchers agree that having linguistic information available while building datasets results in more homogeneous datasets (Evert and Krenn, 2001),

and therefore it is expected that the statistical models based on these will be less sensitive to the noise in the data. When available, researchers have made use of large human-annotated treebanks,[5] since they constitute the highest quality extraction data a computational linguist can have. When human-annotated data is not available, annotation tools are used. One of the risks of adopting a hybrid approach is that errors in annotation may contribute spurious outcomes or candidates to datasets.

This dissertation presents a hybrid approach tested on the task of identifying Dutch *collocational prepositional phrases* and *support verb constructions*. In the remainder of this chapter we describe previous attempts at the identification of collocations and multi-word lexemes using a hybrid approach. The acquisition task involves a sequence of steps: pre-processing extraction data, applying statistical tests and evaluation.

3.2 Pre-processing extraction data

Pre-processing involves preparation of the extraction data by adding some linguistic information. The amount of linguistic information primarily depends on available resources for the specific language, but also on the structural complexity of the collocations to be extracted. For example, if one aims at identifying collocations made up of two adjacent nouns such as *traffic light*, a corpus annotated with only part of speech tags may be good enough. By contrast, if one wants to identify collocations made up of a verb and its direct object such as *make headway*, a corpus annotated with grammatical functions may facilitate the collection of more relevant candidates.

3.2.1 Extraction data

Most current work on collocation identification in English performs extraction of datasets from fully parsed data (Lin, 1999; Blaheta and Johnson, 2001; Pearce, 2002). Some approaches rely on a large treebank from where candidate tuples are selected on the basis of lexical and morpho-syntactic constraints. When a large treebank is not available, extraction of datasets can be done from automatically annotated data, such as part-of-speech-tagged corpora (Church and Hanks, 1990; Smadja, 1993; Dias et al., 1999), chunked data (Krenn, 2000a; Kermes and Heid, 2003) or the output of a lexicalized stochastic grammar model (Lin, 1998; Zinsmeister and Heid, 2002; Schulte im Walde, 2003).

[5] A treebank is a collection of hand-corrected sentences fully annotated with syntactic structure and dependency relations.

3.2.2 Annotation tools for Dutch

In the course of this project, the `Alpino` parser, its lexicon and grammar continue being expanded and improved.

`Alpino` is a wide-coverage parser for Dutch informed by a constraint-based grammar (van der Beek et al., 2002b). Based on a lexicalist constraint-based grammar framework (Head-Driven Phrase Structure Grammar) (Pollard and Sag, 1994), the Alpino grammar incorporates lexical information that encodes properties concerning part of speech, morphology, subcategorization requirements and dependency relations. The grammar currently licenses a variety of syntactic constructions like subordinate clauses, (indirect) questions, imperatives, (free) relative clauses, comparatives, a wide range of verbal and nominal complementation and modification patterns, verbal crossing-dependency constructions, extraposition, and coordination (van der Beek et al., 2002c). The lexicon contains approximately 47,000 lemmas. Lemmas specify (if applicable) subcategorization frames enriched with dependency relations and some lexical restrictions. More than 70 different subcategorization frames specify the complementation requirements of verbs, nouns and adjectives.

Among the parser's various components there is a highly accurate POS-tagger and a maximum entropy disambiguation module that boost the reliability of the parsed output. Currently, the POS-tagger reaches 95% per tag average accuracy while using a very large tagset (Prins and van Noord, 2004). The parser reaches about 87% per sentence average concept accuracy.[6] To be able to repeatedly test and evaluate the impact that changes made to the lexicon, the grammar or to some of the other parser's modules have on the performance of the parser, a treebank has been created. The *Alpino Treebank* includes more than 7,000 sentences extracted from written text that were annotated semi-automatically and hand-corrected (van der Beek et al., 2002a).

3.2.3 Sampling and representation

Corpus patterns whose component words satisfy certain constraints (word category, dependency relation, syntactic relation, etc.) are collected into a dataset. Once a pattern is selected, all instances of that pattern present in the extraction data should be equally likely to be selected into the dataset. If the statistical scores are to be reliable, the selection from the corpus should be a random sampling, such that each object (pattern) in the data sample is selected via independent and identical trials (Pedersen, 1996).

[6] Concept accuracy reflects the percentage of named dependency relations within a sentence that the parser got correct.

Sampling approaches differ depending on (i) the available linguistic annotation and (ii) the type of the target multi-word lexemes. Fully-parsed extraction data allows the researcher to specify more morpho-syntactic constraints than part-of-speech tagged data.

Using part-of-speech tagged data, Church and Hanks (1990) extract *verb particle* outcomes from their corpus and Daille (1996) extracts potential terms from a technical French corpus. Krenn (2000b) describes an experiment to extract support verb constructions and figurative expressions that consist of a VERB and a PP from partially chunked data. Phrasal chunks and clause boundary information allow Kermes and Heid (2003) to collect *adjective verb* collocations from a German corpus. If a syntactic or dependency relation ought to hold between the candidate's component words, fully parsed data is necessary. Thus, Zinsmeister and Heid (2003) use an automatically parsed newspaper text corpus to identify *adjective noun verb* collocational or idiomatic triples in German. Schulte im Walde (2003) uses parsed data to build a database of different collocation types in German.

The independent and identical trial assumption is often violated or ignored. Heuristics and filtering strategies are applied to filter *a priori* judged 'noise' in the datasets before applying statistical tests. Filtering strategies make use of a window technique, stop lists, etc. It remains necessary to track the impact of these filtering techniques on the identification process results.

It is important to mention that datasets extracted from automatically annotated corpora may contain errors. Because of tagger or parser errors, spurious candidates end up in a dataset or existing patterns in the corpus are absent in the dataset. Although spurious outcomes add noise, Evert and Kermes (2003) argued that these errors mostly affect low-frequency types for which the statistical measures are not reliable. Nevertheless, many interesting collocations and multi-word lexemes exist among low-frequency types, therefore one should look out for and beware of systematic annotation errors.

Outcomes are represented as *tuples*. A *tuple* may consist of word forms, lemmas or a combination of both. Researchers report that for some types of collocation, using lemmas gives better results than using word forms (Krenn, 2000b). The underlying reason for using lemmas is to cluster the various word forms whose individual mass distributions may not be statistically significant. In addition to lemmas, Lin (1998) employs the dependency relation between two lemmas ($head\_verb, dependency\_relation, head\_noun$) in order to identify various types of collocations (verb-object, subject-verb, adjective-noun, etc.). Schulte im Walde (2003) follows Lin (1998) and also includes the dependency relation in the tuples.

3.3 Collocation statistics

A dependence between two random variables can be identified with various statistics. Once the probability distribution of each random variable is known, a statistical test determines how similar or dissimilar the probability distributions of the random variables in question are.

Words may instantiate possible values of categorical (random) variables. The 'lexical affinity' between two words can be thought of as the lexical association between the two words and it can be measured as the statistical dependence between two random variables.

To represent the data in terms of a statistical model the components of each candidate bigram, that is, its unigrams, are represented by the binary variables X and Y. Each bigram will have one of four possible classifications corresponding to the possible combinations of these variable values. The possible values of the 2 random variables are cross-classified, the frequency of each possible classification in the sample is collected, and these are represented in a 2 by 2 table, known as a CONTINGENCY TABLE.

3.3.1 Contingency table

Table 3.1 shows the frequency count data that summarizes the hypothetical frequency distribution of a bigram (X, Y) where the random variables have the values $(X = houd, Y = in\_gaten)$. This contingency table cross-classifies the frequency distribution of the two variables X and Y. Rows correspond to the distribution of the X variable that may take two values, either $X =$ *houd* or $X = \neg$ *houd*. Columns give the distribution of the variable Y, where either $Y = in\_gaten$ or $Y = \neg$ *in_gaten*. $\neg$ in front of a unigram stands for 'NOT' and represents the complement set of that unigram.

	w_2 in_gaten	$\neg\ w_2$ $\neg$ in_gaten	Row marginals
w_1 houd	O_{11} 125	O_{12} 7785	O_{1+} 7910
$\neg\ w_1$ $\neg$ houd	O_{21} 213	O_{22} 1159283	O_{2+} 1159496
Column marginals	O_{+1} =338	O_{+2}= 1167068	N = 1167406

Table 3.1: Contingency table for the bigram (*houd*, *in_gaten*), representing the lexicalized expression *iets in de gaten houden* 'keep an eye on sth'.

We refer to the frequency counts for a given bigram (w_i, w_j) in our dataset as the observed frequency O_{ij}. Index i ranges over those words that may occur as first word (w_1) in a bigram and, index j applies to those words seen as second word (w_2). Thus, the O_{11} value gives the raw frequency of the bigram (*houd*, *in_gaten*) in the dataset (that is, 125). O_{1+} and O_{2+} give the row marginal frequencies; these marginal counts correspond to the number of times w_1 and $\neg w_1$ are found in position 1 in a bigram. The column marginals O_{+1} and O_{+2} give the total frequencies for w_2 and $\neg w_2$ as the second unigram in a bigram. The sum of all observed frequencies O_{ij} gives the total sample size N.

Row marginals, column marginals and N (sample size) are needed to calculate the expected frequency (E_{ij}) values for each cell in the contingency table. The expected frequency of a unigram indicates how likely it is to observe the unigram in a given population by chance. This is expressed in equation 3.1. The expected frequency of a bigram (w_i, w_j) reflects how likely w_i is to be seen next to w_j by chance. Equation 3.2 gives the expected frequency of a bigram i.e. the probability of seeing its component unigrams together if they always occur independently of each other. The expected frequency of a cell is the product of the corresponding row and column marginals divided by the sample size. Several statistics require expected frequencies.[7]

$$E(w_i) = \frac{O(w_i)}{N} = P(w_i) \tag{3.1}$$

$$E(w_i, w_j) = E(w_i)E(w_j) \tag{3.2}$$

3.3.2 Association measures

Statistical tests commonly used in identifying word associations or collocations are often referred to as association measures (AMs) (Krenn, 2000a). In this section, we describe a few association measures, the required data and the interpretation of the final score. For each measure, we show an example of how the score is computed based on the contingency table in Table 3.1. For further details about the statistical tests, the reader is referred to Agresti (2002), Evert (2004) and Oakes (1998).

Mutual information Motivated by information theory, mutual information measures how much information a variable tells us about another variable. Church and Hanks (1990) call this measure the *association ratio*. The

[7]The expected frequencies of the contingency table in Table 3.1 are: E_{11}=2.29, E_{12}=7907.7, E_{21}=335.7 and E_{22}=1159160.

mutual information score for the bigram (w_i, w_j) is computed as the logarithm of the probability of seeing two words w_i and w_j together, divided by the product of the words individual probabilities (whether they occur together or in isolation):

$$I(w_i, w_j) = \log_2 \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \tag{3.3}$$

A large mutual information score reflects the circumstance that a given bigram is found more often than chance, i.e. the probability of seeing a bigram is greater than the probability of seeing it if the two words in the bigram were in fact independent. In this case, a large mutual information score rules out the null hypothesis of independence between w_i and w_j.

Mutual information of the bigram $(houd, in\_gaten)$ is computed as follows:

$$\begin{aligned} I(houd, in\_gaten) = \log_2 \frac{P(houd, in\_gaten)}{P(houd)P(in\_gaten)} = \log_2 \frac{\frac{125}{1167406}}{\frac{7910}{1167406}\frac{338}{1167406}} = \\ \log_2 \frac{0.0001}{1.96e-06} = log_2 54.58 = 5.77 \end{aligned} \tag{3.4}$$

In practice, mutual information may misclassify some low-frequency data as collocations. To illustrate this, we compute the corresponding score for the bigram $(houd, met\_popmuziek)$ that occurs only once in the dataset:

$$\begin{aligned} I(houd, met\_popmuziek) = \log_2 \frac{P(houd, met\_popmuziek)}{P(houd)P(met\_popmuziek)} = \\ \log_2 \frac{\frac{1}{1167406}}{\frac{7910}{1167406}\frac{1}{1167406}} = \log_2 \frac{8.56e-07}{7.33e-13} = log_2 \frac{1167406}{7910} = 20.15 \end{aligned} \tag{3.5}$$

The mutual information score assigned to $(houd, met\_popmuziek)$ (20.15) is much larger than the score assigned to $(houd, in\_gaten)$ (5.77). The former score wrongly predicts an association between *houd* and *met_popmuziek*. There is only one occurrence of *met_popmuziek* in the dataset next to *houd*. If this were a representative reflection of its frequency, then the calculation above would not be incorrect.

Pearson's χ^2 The χ^2 metric computes for a bigram (w_i, w_j) how much the observed frequency of each cell of the table in Table 3.1, such that,

$$(w_i, w_j) \in \{\ w_1w_2, w_1\neg w_2, \neg w_1w_2, \neg w_1\neg w_2\ \}$$

deviates from the expected frequency.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.6}$$

Greater differences between the observed O_{ij} and the expected E_{ij} frequencies produce larger χ^2 values, therefore stronger evidence to reject the null hypothesis (H_0) that no association holds between the words in the bigram.

We compute the χ^2 score for our running example as follows:

$$\begin{aligned}\chi^2 = (&\frac{(125 - 2.29)^2}{2.29} + \frac{(7785 - 7907.7)^2}{7907.7} + \frac{(213 - 335.7)^2}{335.7} \\ &+ \frac{(1159283 - 1159160)^2}{1159160}) = 6621.63\end{aligned} \tag{3.7}$$

The χ^2 score for the low-frequency bigram $(houd, met\_popmuziek)$ is 146.58 – much smaller than the score assigned to true collocations. χ^2 is less sensitive to very low frequencies. Intuitively, this is due to the fact that χ^2 is based on observed frequencies and not merely relative frequencies, which mutual information relies on.

Log-likelihood ratio This measure works well with both large and small sample sizes and furthermore, it allows a direct comparison of the significance of common and rare phenomena (Dunning, 1993). This ratio can more easily be computed by the log-likelihood chi square ratio (aka. G^2) (Evert, 2002):

$$G^2 = 2 \sum_{i,j} O_{ij} \log_2 \frac{O_{ij}}{E_{ij}}$$

For the bigrams (w_i, w_j), G^2 adds up the product of the observed value O_{ij} and the logarithm result of dividing the observed frequency O_{ij} by the expected frequency E_{ij}. The larger the G^2 value the more evidence the model has to reject the null hypothesis. Again, large scores are interpreted as an association between the words in a bigram.

We take up our example again and compute the G^2 score in the following way:

$$\begin{aligned}G^2 = 2((&125 \log_2 \frac{125}{2.29}) + (7785 \log_2 \frac{7785}{7907.71}) + (213 \log_2 \frac{213}{335.7}) \\ &+ (1159283 \log_2 \frac{1159283}{1159160})) = 808.03\end{aligned} \tag{3.8}$$

The low-frequency bigram (*houd*, *met_popmuziek*) is assigned a score of 9.98, certainly much smaller than the score assigned to true collocations. We attribute this to G^2's incorporation of concrete frequencies.

Dunning (1993) argues that when the event space (number of outcomes) is large enough, Pearson's χ^2 and the log-likelihood ratio give similar results; however, when the observations space is small, Pearson's χ^2 overestimates the significance of rare data. The corresponding G^2 and χ^2 scores attributed to the bigram (*houd*, *met_popmuziek*) are rather different; this suggests that the data is insufficient for the χ^2 score to be reliable.

Fisher's exact test This test may be applied to small-sample distributions as well as large-sample distributions. Fisher's test computes the hypergeometric probability of observing frequency O_{11} in a contingency table based on the sample size (N), the row marginal total O_{1+} and column marginal total O_{+1}. Computing the hypergeometric probability $P(O_{11})$ requires the computation of the probability of observing O_{11} in all possible contingency tables that have the same sample size (N), row marginal total O_{1+} and column marginal total O_{+1}.

$$P(O_{11}) = \frac{\binom{O_{1+}}{O_{11}}\binom{O_{2+}}{O_{+1}-O_{11}}}{\binom{N}{O_{+1}}} \tag{3.9}$$

Pedersen (1996) recommends this test for small samples only, because it requires heavy computation. We use Pedersen's implementation of a left-sided Fisher's test. This version of the test computes the probability of observing the same or a smaller value for O_{11} in the contingency tables with same N value and row and column marginals. A high left-sided probability indicates that one is very unlikely to observe the given bigram more frequently than one already has, thus an indication that these pair of words is somehow special.

Both bigrams (*houd*, *in_gaten*) and (*houd*, *met_popmuziek*) are assigned a left Fisher value of 1.0, the maximum value.

Salience This measure is an adjustment to pointwise mutual information and estimated as the product of mutual information and log frequency:

$$I(w_i, w_j) \log_2 O_{ij} = \log_2 \frac{P(w_i, w_j)}{P(w_i)P(w_j)} \log_2 O_{ij} \tag{3.10}$$

The test was originally proposed by Lin (1998), and later used for collocation extraction in lexicography by Kilgarrif and Tugwell (2001). Kilgarrif

and Tugwell (2001) use salience to prevent low-frequency candidates from ending among the higher-ranked candidates.

The salience score of the bigram $(houd, in\_gaten)$ is:

$$salience(houd, in\_gaten) = \log_2 \frac{P(houd, in\_gaten)}{P(houd)P(in\_gaten)} \log_2 O(houd, in\_gaten) = \log_2 \frac{\frac{125}{1167406}}{\frac{7910}{1167406}\frac{338}{1167406}} \log_2 125 = \log_2 \frac{0.0001}{1.96e-06} * 6.96 = log_2 54.58 * 6.96 = 40.19 \tag{3.11}$$

The corresponding value for hapaxlegomena like $(houd, met\_popmuziek)$ is 0. The salience score of hapaxes will always be null, because the $log_2(1) = 0$.

3.3.3 Hypothesis testing

Hypothesis testing helps the researcher make a categorical decision. Statistic tests used in collocation identification inform the decision whether a candidate in the sampled data is a collocation or not. These tests measure the divergence between the probability distribution of the candidate's observed frequency in the data sample and the probability distribution of the candidate's expected frequency under the assumption that the candidate's component words are independent of each other. In short, two hypotheses are compared: the *null hypothesis* (H_0) and the *alternative hypothesis* (H_a). The *alternative hypothesis* (H_a) states that the probability distributions of the component words exhibit a dependence, thus, suggesting that the candidate is a collocation. In contrast, the *null hypothesis* (H_0) states that the probability distributions of the component words are independent.

The AMs are applied to the contingency table that summarizes the frequency distributions of the candidate ngram and its component unigrams. If the resulting AM score assigned to the candidate ngram is *significantly* different from a hypothesized score (often 0), the null hypothesis may be rejected.

One of the challenges faced in identification of collocations and multiword lexemes is to find a statistic test that among all candidates assigns a high score to those whose composite words show a strong dependence; in other words, one aims at finding candidates whose statistic score suggests we can confidently reject the *null hypothesis*.

3.4 Loglinear models

Both a loglinear model and the association measures infer associations between variables. The structural form of a loglinear model models the patterns of association and interaction between the variables. Thus, the advantage of using a loglinear model is that this type of models can handle more complex relations such as analyzing simultaneously the effect of several variables (Agresti, 2002).

The problem of computing the association between various words can be decomposed into various measurements: on the one hand, assessing how frequent the word combination (as a whole) is in the language and on the other hand, measuring the partial associations between every two words in the expression allowing for the possibility of accidental word co-occurrences.

Let us now consider an example. The association score between a verb, a preposition and a noun (e.g. *take into account*) can be approximated by measuring the binary associations between the verb and the preposition, the verb and the noun and, the preposition and the noun. However, it is possible that a verb frequently co-occurs with a preposition even though the two words bear no lexical relation. Such combinations result from the fact that two (highly) frequent words or phrases cooccur in the same context by chance; they are called *anticollocations* and every identification model should beware of them. To minimize the chance of anticollocations, the frequency of the individual words (i.e. the verb, preposition and noun) negatively affects the association score of the whole word combination.

Given a three-dimensional contingency table (one more dimension than Table 3.1) that cross-classifies the possible values of three categorical variables X, Y and Z, the most general loglinear model where i ranges over the different values of X, j ranges over the different values of Y, and k ranges over the different values of Z has the form:

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

This model states that the expected frequencies μ_{ijk} of the observations O_{ijk}, can be predicted on the basis of the total of observations (λ) and a number of effects λ_i^X, λ_j^Y and λ_i^Z. The λ terms are model parameters and their values are estimated using maximum likelihood estimators. This is the simplest model that should hold if the three variables are independent. If there is reason to believe that the three variables X, Y and Z are not truly independent, a more complex model should add term effects that measure the interaction between every two variables and also the interaction between the three variables. The SATURATED loglinear model adds such interaction

terms namely, λ_{ij}^{XY}, λ_{jk}^{YZ}, λ_{ik}^{XZ} and λ_{ijk}^{XYZ}. The model has the form:

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ} \tag{3.12}$$

Applied to language models, categorical variables have nominal values, that is, words. The models can be used to identify interactions between words that can be part of a collocation. Indirectly the models are applied to identify expressions made up of words that exhibit a strong interaction between them.

Blaheta and Johnson (2001) apply a *saturated* loglinear model to identify English multi-word verbs (e.g. *look forward to*). The model parameters measure and consequently, reflect the degree of association between the composite words (verb and particles) in a candidate ntuple. They use maximum likelihood estimators for the n-way interaction terms in loglinear models (if n=3, the interaction term would be λ_{ijk}^{XYZ} above). These interaction terms measure how strong the association between a verb and the accompanying particles inside an ntuple is. Refer to Blaheta and Johnson (2001) for details on how the interaction parameters are actually computed. Blaheta and Johnson (2001) report good results when the model is used to identify multi-word verbs in English. Furthermore, the loglinear model applies to tuples of any length and the model does not suffer from the low-frequency bias.

According to the loglinear model, the significant interaction term assigned to the example *trigram* $(houd, in, gaten)$ is 11.1163. The loglinear is designed to control for the low-frequency bias; a frequency cutoff and/or a significance value discard hapax legomena such as $(houd, met, popmuziek)$. Contrary to the association measures, this model is applied directly on trigrams.

3.5 Evaluation methodology

A standard evaluation methodology has not yet been proposed, despite the rigorous method comparison proposed by Evert and Krenn (2001). The lack of a consistent evaluation methodology is a by-product of the lack of any agreed-upon definition of collocation (Pearce, 2002). Nonetheless, a consensus exists that an identification method is successful if it assigns the higher scores to true collocations and the lower scores to uninteresting combinations. In addition, the method ought to identify as many collocations as there are in the extraction corpus. In order to decide whether an expression in the output list is a true collocation, researchers have adopted one of the following approaches:

- compare the ranked list of extracted collocation candidates to a list of expressions compiled from existing lexical resources (dictionaries, glossaries, index, term bank, etc.) or collocations that have been annotated in a (corrected) treebank;

- native speakers assess the degree of collocativity of the extracted collocation candidates; or

- task-based evaluation: test how the addition of the extracted collocations in a lexical resource improves the results of a system in a given application (e.g. parsing).

3.5.1 Validation data

Machine readable dictionaries, such as specialized dictionaries of idiomatic expressions or monolingual dictionaries for language learners have been used to compile a list of true collocations (or other multi-word expressions) (Lin, 1999; Pearce, 2002; Schulte im Walde, 2003); such validation data list is known as the GOLD STANDARD. Besides dictionaries, a term bank (Lin, 1998) and a term index included in a book (Arehart, 2003) have also proved to be useful.

For various reasons, compilation of a large gold standard list from existing dictionaries might not be possible. In this case, a random sample of words (nouns, verbs) is selected. Then, all collocations with those words are (manually) extracted from a dictionary. This 'dictionary list' is the gold standard. The ranked list of candidate collocations is checked against the gold standard. Lin (1998) applies this methodology.

Machine readable dictionaries exist for many languages, but sometimes the researcher has no access to them. Furthermore, the expressions one is interested in are not consistently annotated in published dictionaries. In this case, the researcher opts for a different approach by either asking a lexicographer (Smadja, 1993) or native speakers (Blaheta and Johnson, 2001) to judge the validity of the retrieved expressions. Alternatively, the researchers themselves search for true collocations in the candidate dataset (Krenn, 2000b; Evert and Krenn, 2001) thus, manually compiling a gold standard list. Whereas a skilled lexicographer may not hesitate about what a collocation is, nonexpert human judges or the researchers themselves typically find the task rather hard, thus they disagree on their judgements. This obviously adds more difficulties during the manual compilation of a gold standard list.

3.5.2 Quantitative evaluation

Precision and recall are typically used to account for the coverage of statistical models. Precision measures the proportion of selected items that are correct, thus, it measures the quality of the retrieved material. Alternatively, recall measures the proportion of correct items that were selected, i.e. the effectiveness of the system.

$$Precision = \frac{correct\_items\_selected}{correct\_items\_selected + incorrect\_items}$$

$$Recall = \frac{correct\_items\_selected}{correct\_items}$$

Krenn (2000b) and Evert and Krenn (2001) evaluate the performance of the models on the basis of (i) varying frequency thresholds applied on datasets and (ii) varying N-best candidates lists. Evert and Krenn (2001) argue that precision and recall should be assessed for data from different frequency ranges and for data among the high and low scores to ensure that results are not a matter of chance. Thus, they plot recall and precision graphs for the whole set of candidate data.

Standard precision and recall can only be computed if all true collocations have been identified in the dataset. If all true collocations have not been identified, *approximate* precision and recall can be computed on the basis of a reference list of collocations. The evaluation described in chapter 4 reports approximate precision given that a rather large list of collocational prepositional phrases is available. Instead, in chapter 5, we introduce a different evaluation metric known as *uninterpolated average precision* (Manning and Schütze, 1999).

3.6 Choosing the best approach

Various investigations target different collocation and multi-word lexeme types and they use different corpora, annotation tools, sampling strategies, statistic tests and evaluation methodology. There is no best approach that one can adopt in identification of collocations and multi-word lexemes.

In general, researchers prefer to use richly annotated data, thus chunked and shallow- or fully-parsed data are most frequently used.

Among the statistical tests described, researchers' opinions vary and sometimes contradict each other. There seems to be a consensus in that

the mutual information score overestimates the association between words in low-frequency candidates (Church and Hanks, 1990; Dunning, 1993; Evert and Krenn, 2001). Dunning (1993) argues that when the event space (number of outcomes in dataset) is large enough, Pearson's χ^2 and the log-likelihood ratio give similar results; however, experiments with natural language show that the two tests give very different results. When the observations space is small, Pearson's χ^2 overestimates the significance of rare data. Contradicting Dunning (1993), Pedersen et al. (1996) show that G^2 and Pearson's χ^2 scores are unreliable when: (i) the sample is not large enough, or (ii) the number of cells in the contingency table is larger than the sample size or (iii) a cell's expected frequency is less or equal than 5. When one suspects the 3 conditions mentioned above are not met, Pedersen et al. (1996) recommend using an exact distribution of a goodness of fit test statistic. Two ways are possible: (i) enumerate all elements of that distribution as in Fisher's exact test, or (ii) sample from that distribution using a Monte Carlo sampling scheme.

Systematic comparison of association measures reveals differences in precision and recall depending on the collocation type. Evert and Krenn (2001) observe that whereas the log-likelihood test and the t-Test[8] are good measures to extract adjective-noun pairs from a German newspaper corpus, the t-Test performs better in identifying `preposition + noun + verb` constructions. If the dataset includes low frequency data, Evert and Krenn (2001) showed that the log-likelihood test reaches higher precision than the t-Test, χ^2 and mutual information. If low frequency data is discarded, the performance of t-Test, log-likelihood and raw frequency is rather similar when extracting `preposition + noun + verb` collocations. In a different comparison of association measures, Krenn and Evert (2001) investigate what measures are more useful to identify PP-verb collocations and conclude that mere co-occurrence frequency achieves significantly better results than the AMs. Judging from the results reported in previous work, the log-likelihood ratio seems to be the most reliable and in fact, the preferred test.

One should mention that in collocation identification a large percentage of the candidates constitute hapax legomena (one occurrence candidates) and dislegomena (two occurrence candidates). Often 70% of the extracted dataset is made up of hapax legomena (Zinsmeister and Heid (2003) describes such a typical scenario). The success of a method relies on how accurately the statistic identifies true collocations among low-frequency data.

Previous work tried to tackle the low-frequency bias by applying a frequency cutoff that discards those candidates in datasets whose frequency is below a chosen threshold. A frequency threshold of $f \geq 5$ is the most

[8]Refer to Evert and Krenn (2001) for a description of the test.

commonly applied (Evert and Krenn (2001); Pearce (2002); Zinsmeister and Heid (2003) among others). The cutoff is not always the best solution because it leads to the loss of the infrequent collocations. On the other hand, with more data the loss of collocations might be lessened. The cutoff value needs to be empirically established, because it varies with corpus size and collocation type.

The association measures have been mostly applied on bigrams. When the collocation candidates consist of three words, triples are treated as bigrams. As examples, Krenn (2000b) and Zinsmeister and Heid (2003) applied the tests to candidate triples treated as $(verb, prep\_noun)$ and $(adj\_noun, verb)$ bigrams, respectively. Measures generalized to ngrams when $n \geq 3$ have been applied by Blaheta and Johnson (2001) (a log-linear model) and by Dias et al. (1999) and Kaalep and Muischnek (2003) (mutual expectancy).

Previous research agreed that addition of linguistic information improves the accuracy of statistical models in identifying lexical associations. There has been much progress in improving the quality of the retrieved expressions, however the identification problem is not solved yet. The main limitations of the existing models pertain to the length of the candidate ngram, the low frequency data (a non-negligable proportion of potential candidate expressions) and the lack of validation data.